Validation of Software Releases for CMS

Oliver Gutsche[1] **on behalf of the CMS Computing and Offline Projects**

[1] Fermi National Accelerator Laboratory, CMS/CD MS 205, P.O.Box 500, Batavia, IL, 60510, USA

E-mail: `gutsche@fnal.gov`

Abstract. The CMS software stack currently consists of more than 2 Million lines of code developed by over 250 authors with a new version being released every week. CMS has setup a validation process for quality assurance which enables the developers to compare the performance of a release to previous releases and references.

The validation process provides the developers with reconstructed datasets of real data and MC samples. The samples span the whole range of detector effects and important physics signatures to benchmark the performance of the software. They are used to investigate interdependency effects of all CMS software components and to find and fix bugs.

The release validation process described here is an integral part of CMS software development and contributes significantly to ensure stable production and analysis. It represents a sizable contribution to the overall MC production of CMS. Its success emphasizes the importance of a streamlined release validation process for projects with a large code basis and significant number of developers and can function as a model for future projects.

1. Introduction

The Large Hadron Collider (LHC) at CERN, Geneva, Switzerland [1] is expected to record the first proton-proton collisions in September 2009. The Compact Muon Solenoid experiment (CMS)[2], one of the two startup experiments of the LHC, is ready to start collecting data and eagerly awaiting the first collisions. A world-wide distributed computing infrastructure [3] will be used to process the data, to simulate collisions and to analyze both to extract physics results. The binding element between all these activities is the CMS software stack [4]. It contains the persistency formats to store CMS' data in files on disk and tape and the software for data recording, reconstruction and analysis. The stack also contains all necessary components to perform Monte Carlo simulations (MC) using different generators.

The CMS software consists of more than 2 Million lines of code not including external third party packages and has a very active developer community of more than 250 individual developers [5]. To facilitate development, CMS consolidates its code base regularly into releases. Different releases are grouped into release cycles dedicated to a specific purpose (e.g. data taking and reconstruction in 2009/2010 LHC collision period, integration of a new ROOT version, etc.) to aggregate specific feature sets of the software stack. There are 3 main types of releases:

Pre-Release: consolidates the current state of the code base at a given point in time. It is primarily a test release to present the developer with a consistent snapshot of the development status and allows to test interdependencies between different software components developed in parallel. Pre-Releases are not used for official tasks like data taking and reconstruction, MC production or analysis.

Release: closes a development cycle. It contains the final feature set of the respective cycle. New features are not allowed to be integrated anymore and have to go into the next release cycle. The release is used for official central activities and analysis. A release is distributed and installed world-wide on all levels of the CMS tiered computing infrastructure [3].

Amendment Release: only integrates bug fixes to an earlier release to fix specific problems. It supersedes the previous release and is used for official activities and distributed world-wide.

CMS is undergoing a very rapid development cycle in anticipation of first collision data. On average, a new pre-release is provided to the developers and the collaboration every week. In 2008, CMS released 78 different releases of the software stack in 4 different release cycles (see Fig. 1).

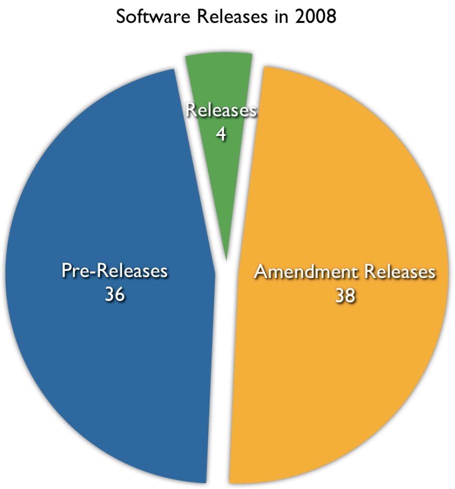

Figure 1. Distribution of release type of the software releases published by CMS in 2008.

The rapid development cycle of CMS and the resulting high number of releases requires a thorough quality assurance process. To guarantee stable and performant releases while supporting a high turn around and diverse development, CMS implemented an advanced central release validation process.

2. Release Validation

CMS implemented a central release validation early on to accompany the rapid software development. It is now an integral part of the complete software development process and has an important purpose to fulfill:

- Release validation guarantees that all software components of a release work together without interference or failures during execution.
- It checks that a release is compatible with the global production and processing infrastructure of CMS [6].
- Release validation is used to validate the correctness of the produced physics output and performance of a release in terms of:
  - Algorithmic performance,
  - Stability at larger scales (e.g. number of events),

– Memory and time consumption

In general, release validation produces reference MC simulation samples and reconstruction passes of detector data for every release. The samples are provided to the developers and the collaboration promptly for the validation of the release. Dedicated resources are used for the production of these samples. They are chosen so that every new release can be installed instantaneously after its announcement and consist of resources at the CERN T0 and the Fermilab T1 center:

CERN: 500 priority batch slots are provided in parallel to the Tier-0 computing resources for a short turn around sample production.

Fermilab: Computing cycles of the 5000 batch slots of the Fermilab Tier-1 computing center are used for release validation in parallel to central production activities like re-reconstruction and skimming. In this mode, Fermilab cannot be used for a short turn around production. It's main usage is to supplement production at CERN by providing with higher statistics samples for validation.

3. Validation Sets

All release validation samples/datasets are grouped into *validation sets*. The composition of the sets is optimized to use the available resources efficiently:

Standard Set: is produced *within 24 hours* at CERN to enable rapid feedback before the next release is published (on average 1 week). The very low latency gives the developers enough time to validate the release and provide fixes for possible problems.

High Statistics Set: is produced *within 1 week* at Fermilab to provide higher statistics comparison samples/datasets to supplement the standard set. The high statistics set is not produced for every pre-release but at least twice per release cycle.

The samples and datasets of the sets span the whole range of detector effects and important physics signatures to benchmark the performance of the software. A summary of the sample composition can be seen in Tab. 1.

Table 1. Composition and statistics of release validation sample sets. Detector data samples have only been added recently and are not mentioned here.

	Full Simulation			**Fast Simulation**		
Generation	Particle Gun	Physics Process	# Events/ Sample	Particle Gun	Physics Process	# Events/ Sample
Standard Set	8	24	9k	6	1	27k
High Statistics Set	12	19	25k	0	8	100k

All samples have been requested by different groups working on software development for CMS to validate their work. The groups can be distinguished by 3 main categories:

- Reconstruction software development,

- Trigger development,
- Alignment & calibration development.

They for example include groups responsible for fast simulation, ECAL lower level reconstruction, track reconstruction and b tagging amongst others.

All groups have defined benchmark samples specifically suited to validate their software responsibilities. These samples differ only in their generator configuration and contain single particle gun samples but also different physics processes like generic QCD to SUSY benchmark points (see Tab. 1).

Resources for the release validation production are limited. Therefore the composition of the sample sets has to be designed carefully to optimize the usefulness of the samples produced and to fulfill all or most of the requests. Where possible, synergies between different sample requests have to be found and exploited. For example, the very high energy background QCD sample for $3000 \leq \hat{p}_T \leq 3500$ GeV is used by the b tagging, the jet reconstruction and the trigger group therefore maximizing the usage and leaving more resources for other samples.

The samples also have to cover different detector conditions. For example, the software has to be validated for the ideally aligned detector and also for the misaligned detector corresponding to the expected alignment accuracy for the startup of data taking. The top quark dataset for example is produced both for ideal and startup conditions. Both conditions are used by the b tagging and HCAL lower level reconstruction groups to validate differences between the conditions whereas the top quark sample in ideal conditions is used by the track reconstruction group and the startup conditions by the trigger group.

Exploiting all possible synergies and overlaps of the different requests enables CMS to produce many different samples and still stay within the resource limits.

4. Validation Set Production

The production workflow of each release validation sample consists of the following components:

(i) Generation of proton-proton collisions for various physics processes or single particles
(ii) Simulation of the detector response
(iii) Local reconstruction of individual detector component signals (RAW detector output)
(iv) Simulation of the trigger response
(v) Reconstruction of objects like tracks, jets and other global event quantities
(vi) Extraction of special event quantities for alignment and calibration (*Al&Ca*) of the detector.

For the full simulation samples, these components cannot be executed in one step. They have to be processed separately in the following steps:

(i) Generation, Simulation, local reconstruction and simulation of the trigger response
(ii) Reconstruction
(iii) Extraction of alignment and calibration event quantities

Each of the steps produces one or more output datasets . The result of step 1 is called the *RAW* dataset and is the input to step 2. The result of step 2 is called the *RECO* dataset and is the input to step 3. Step 3 has several different outputs which contain special data formats for alignment and calibration workflows. They are called the *AL&Ca* datasets (although also step 2 writes out some *Al&Ca* datasets). The release validation sets contain a matrix of different configurations for the different steps combined according to the physics content of the generation.

The release validation workflows are special compared to the normal production workflows used for MC simulation and data taking to simplify the validation for the software developers.

Only the generator information and the RAW event content is written out in production workflows. For the release validation samples, the complete information is written out including the digitized detector signals and special trigger debug information. In addition, all higher level steps are augmented with special validation components storing histograms into the output file. This distorts the performance figures of the release validation workflows (see Sec. 5).

Release validation samples are produced by splitting up the requested statistics in small jobs to efficiently use a batch system. Normally, the different steps would be processed one after the other, finishing the requested statistics for step 1 before starting step 2. This introduces a significant bookkeeping overhead and increases the total processing time because of tail effects. It reduces the efficiency with which the special validation batch queue at CERN is used and the amount of samples which can be produced in the given time window (24 hours for the standard set). CMS optimized the production infrastructure by introducing the concept of *chained processing* (see Fig. 2) because the demand for release validation samples increased significantly over time as more and more components of the software had to be thoroughly tested.

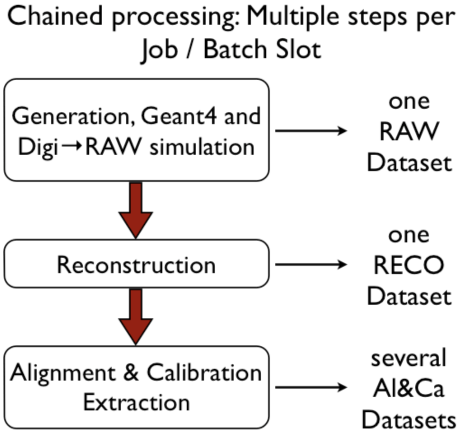
Chained processing: Multiple steps per Job / Batch Slot
Generation, Geant4 and Digi→RAW simulation
one RAW Dataset
Reconstruction
one RECO Dataset
Alignment & Calibration Extraction
several Al&Ca Datasets

Figure 2. Chained processing is an optimization of the CMS production and processing infrastructure to minimize bookkeeping overhead. It runs the steps of the release validation sample production in one job using the output of previous steps directly on the workernode.

With chained processing the production infrastructure is able to process all steps of the production of a release validation sample in one job on one batch slot using the output of a previous step as the input of the next step. This requires that each processing job is able to stage out multiple outputfiles corresponding to the outputs of the different steps. With chained processing, the number of samples could be increased significantly. Fig. 3 shows the special validation queue at CERN during production of several standard release validation sample sets. Each set fits nicely into the 24 hours time window.

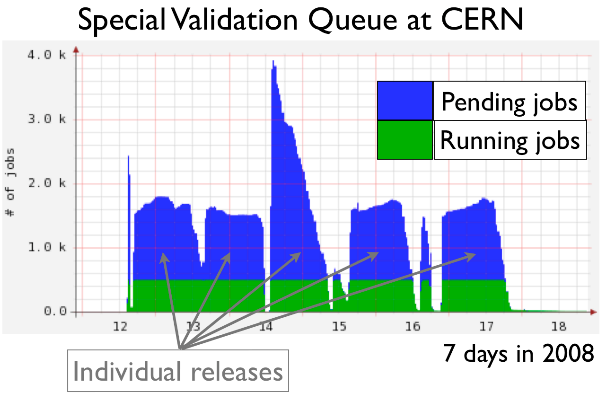

Figure 3. Special validation queue at CERN.

The very efficient usage of processing resources of the chained processing mode for multi-step production workflows was pioneered in release validation and is now also used in CMS' central MC production.

Chained processing enables CMS to produce a significant number of release validation samples with enough statistics to fulfill most of the sample demands by development. Fig. 4 shows the total number of events and Fig. 5 the total number of different datasets for the standard release validation set.

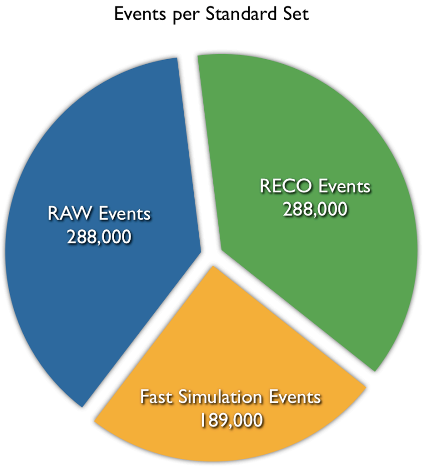

Figure 4. Total number of events for the standard release validation set.

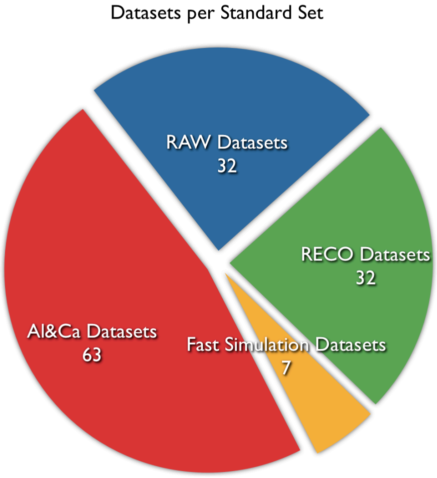

Figure 5. Total number of datasets for the standard release validation set.

The same is shown in Fig. 6 and Fig. 7 for the high statistics release validation set.

Because of the higher statistics and the competition for computing resources with the official production at the Fermilab T1, the high statistics validation set is not produced for every pre-release or release. It is normally produced twice per release cycle.

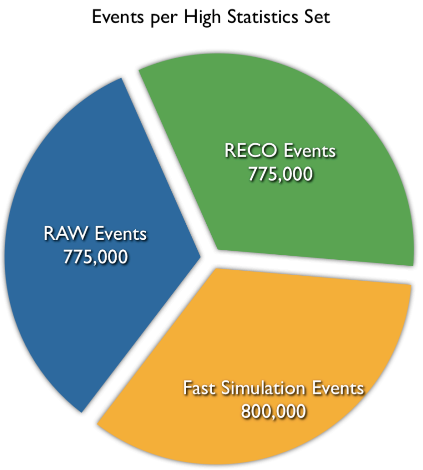

Figure 6. Total number of events for the high statistics release validation set.

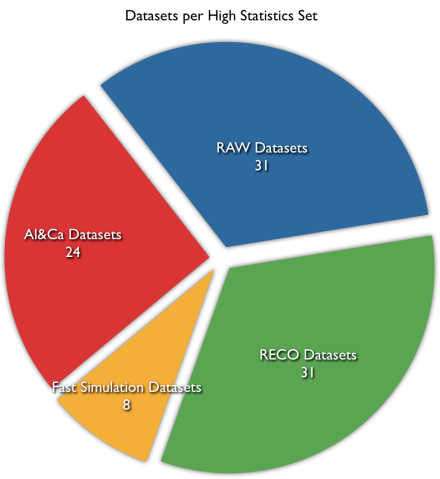

Figure 7. Total number of datasets for the high statistics release validation set.

5. Performance Overview

Global performance numbers are extracted using information from the production jobs themselves from all the release validation samples produced for a certain pre-release or release. These global performance numbers track the maximum virtual memory consumption and the time per event for every step of the release validation sample production. Fig. 8 shows as an example for one release the maximum memory consumption for the simulation step and the time per event for the reconstruction step averaged over all processing jobs per sample.

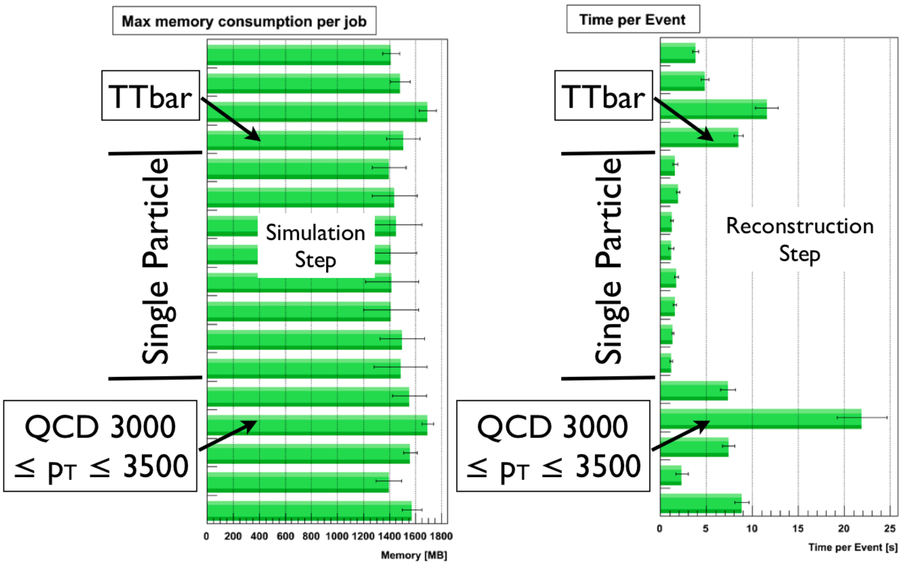

Figure 8. Performance overview from the release validation sample production averaged over all processing jobs of individual samples. *Left:* maximum memory consumption for the simulation step, *Right:* time per events for the reconstruction step.

As mentioned in Sec. 3, these figures present a distorted picture compared to the actual production workflows due to the mentioned changes and addition in the release validation workflows. Still, these numbers can be used to compare the performance within one release and between different releases to see trends in the memory consumption and processing time.

In addition, CMS uses the *performance suite* infrastructure [7] to validate the realistic performance of a release. Here, the exact production workflows are used to avoid distortions due to the special components in the release validation workflows. A set of standardized machines is used to run interactive jobs instead of a batch system which can introduce machine architecture dependencies. Fig. 9 shows the results for the top quark validation sample for an exemplary release.

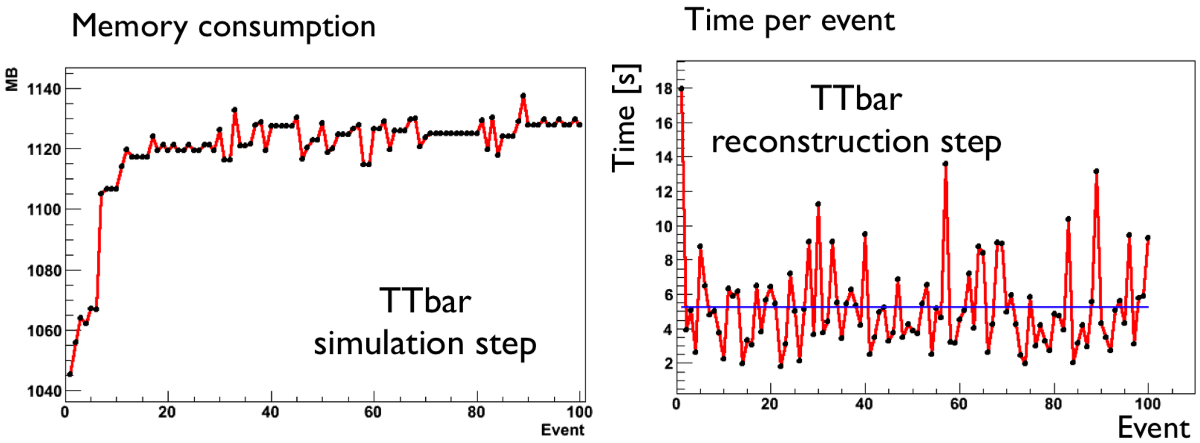

Figure 9. Performance suite result as an example for the top quark validation sample of a release. *Left:* memory consumption for the simulation step per event, *Right:* time per event for the reconstruction step.

6. Release Validation during 2008

Release validation is an integral part of the CMS software development process. In 2008, for validation only CMS produced more than **5100** release validation datasets corresponding to over **114 million events** and **110 terabyte of MC and data samples**. Fig. 10 shows the total number of events produced for release validation in 2008 and Fig. 11 the total data size.

Developers in CMS rely on the timely provision of reference samples to validate their software components. The demand for release validation samples exceeds the production capabilities of the available resources many times. Therefore, compromises in the composition of the release validation sample sets have to be found and synergies have to be exploited where possible. The samples are promptly used after they have been made available to the collaboration. About 2/3 of the release validation samples are used for validation within 2 days of the announcement of the samples, 1/3 is used between 6 and 10 days after the availability of the samples.

7. Conclusions

CMS' release validation process ensures stable software releases for production and analysis. By providing reference MC and data samples for each pre-release and release, developers are able to validate their software components for performance and stability and check interference effects between different components developed in parallel. With the release validation effort, CMS was able to investigate the stability and performance of each release within the very rapid

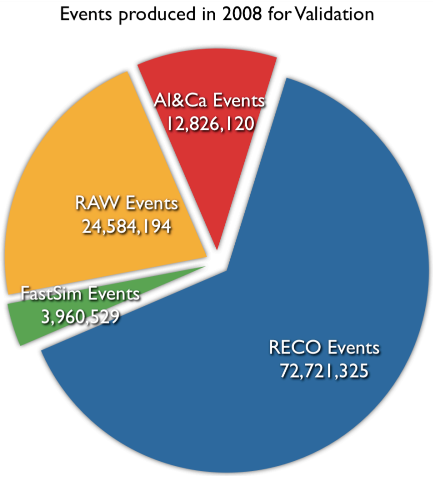

Figure 10. Total number of events produced in 2008 for release validation.

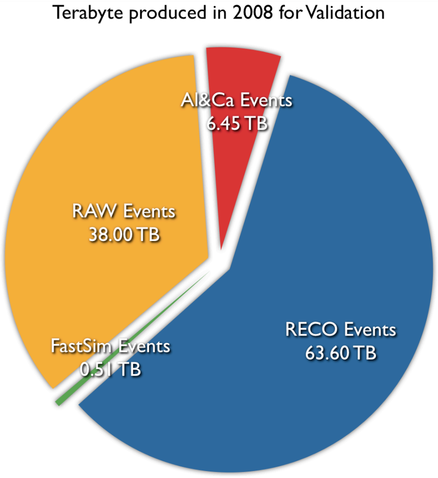

Figure 11. Total amount of terabyte produced in 2008 for release validation.

development cycle of almost one pre-release per week. In 2008, the release validation production was sizeable and resulted in over 114 million events and 110 terabyte of MC and data samples produced. In the future, the release validation sample sets will be adapted further to the needs of the developers and will use detector data samples as input rather than MC generations.

8. Acknowledgments

We would like to thank the funding agencies under which this work was conducted and all CMS collaborators for their help, support and contributions.

References

[1] O. Bruning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole and P. Proudlock, "LHC design report. Vol. I: The LHC main ring", *CERN-2004-003*

[2] R. Adolphi *et al.* [CMS Collaboration], "The CMS experiment at the CERN LHC," JINST **0803**, S08004 (2008) [JINST **3**, S08004 (2008)].

[3] CMS Collaboration, "CMS: The computing project. Technical design report", *CERN-LHCC-2005-023*

[4] C. Jones et al, "The new CMS event data model and framework", Proceedings for Computing in High-Energy Physics (CHEP '06), Mumbai, India, 13 Feb - 17 Feb 2006

[5] P. Elmer, E. Sexton-Kennedy, C. Jones, "The life cycle of HEP offline software", *CHEP '07*

[6] F. van Lingen et. al., "CMS production and processing system - design and experiences", *Poster 82, CHEP '09*

[7] G. Benelli, "The CMSSW benchmarking suite: using HEP code to measure cpu performance", *Poster 69, CHEP '09*